



Statistical Thinking with R

Lei Zhang

Institute of Systems Neuroscience, University Medical Center Hamburg-Eppendorf

Overview

What is your experience with...

- Statistics?
- R? (and / or Matlab?)

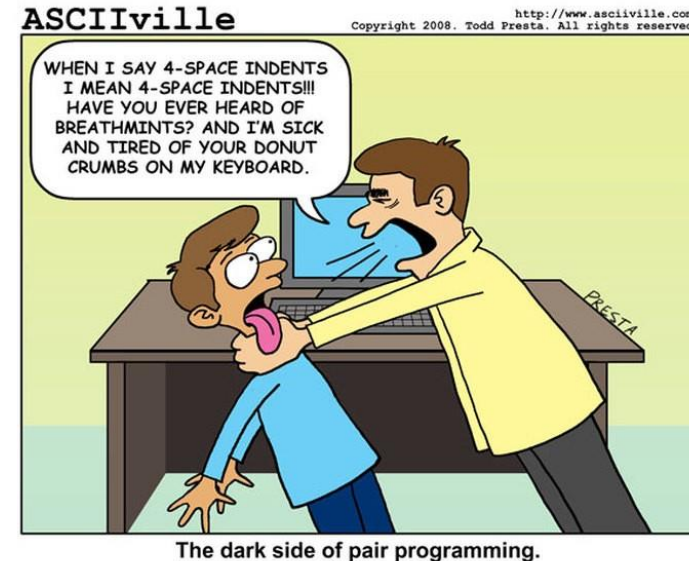
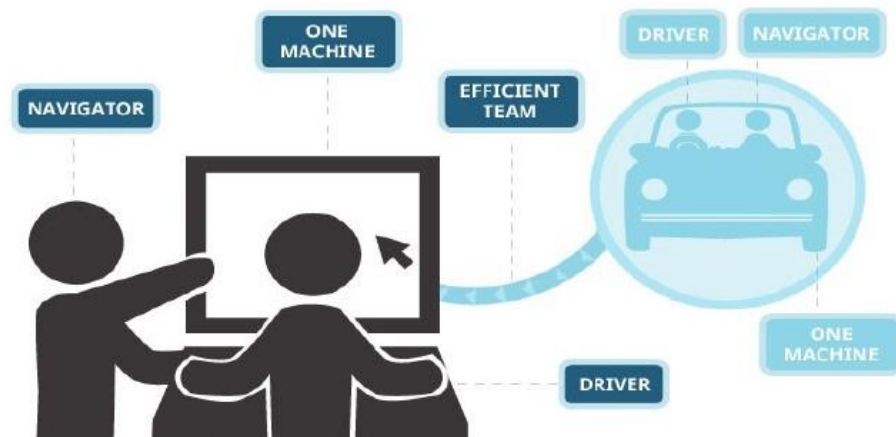
You would like to...

- know how to start and get help with R?
- gain knowledge of (Bayesian) stats?
- start here, know where you can go from it.

How to Get the **Most** out of the Workshop

- Work in pairs: Talk to each other & help each other
- Ask questions
- Try the exercises

PAIR PROGRAMMING



R Basics

- R
 - a programming language for statistical computing
 - R has its own user interface
 - freely available on Windows, Mac, and Linux
- R Studio
 - integrated development environment (IDE) for R
 - a more sophisticated R-friendly editor, with helpful syntax highlight



script editor

The script editor displays the following R code:

```
21 #  
22 library(ggplot2)  
23  
24 myconfig <- theme_bw(base_size = 20) +  
25   theme(panel.grid.major = element_blank(),  
26         panel.grid.minor = element_blank(),  
27         panel.background = element_blank() )  
28  
29 ## normal distribution  
30 # dnorm  
31 g1 <- ggplot(data.frame(x = c(-5, 5)), aes(x)) +  
32   stat_function(fun = dnorm, args = list(mean = 0, sd = 1), size = 3, colour = 'black')  
33 g1 <- g1 + myconfig  
34 print(g1)  
35  
36 # pnorm  
37 g2 <- ggplot(data.frame(x = c(-5, 5)), aes(x)) +  
38   stat_function(fun = pnorm, args = list(mean = 0, sd = 1), size = 3)  
39 g2 <- g2 + myconfig  
40 print(g2)  
41  
42 # qnorm  
43 g3 <- ggplot(data.frame(x = c(0, 1)), aes(x)) +  
44   stat_function(fun = qnorm, args = list(mean = 0, sd = 1), size = 3)  
45 g3 <- g3 + myconfig  
46 print(g3)
```

environment/
command history

console

The console displays the following text:

```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"  
Copyright (C) 2015 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> |
```

file/pkg/img/
etc.

Name	Description	Version
System Library		
abind	Combine Multidimensional Arrays	1.4-3
assertthat	Easy pre and post assertions.	0.1
base64enc	Tools for base64 encoding	0.1-3
BayesFactor	Computation of Bayes Factors for Common Designs	0.912-2
BH	Boost C++ Header Files	1.60.0-1
bitops	Bitwise Operations	1.0-6
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-17
broom	Convert Statistical Analysis Objects into Tidy Data Frames	0.4.1
Cairo	R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output	1.5-9
car	Companion to Applied Regression	2.1-1
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
class	Functions for Classification	7.3-14
cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.3
coda	Output Analysis and Diagnostics for MCMC	0.18-1
codetools	Code Analysis Tools for R	0.2-14
colorspace	Color Space Manipulation	1.2-6
compiler	The R Compiler Package	3.2.3
corplot	Visualization of a correlation matrix	0.73
cubature	Adaptive multivariate integration over hypercubes	1.1-2
curl	A Modern and Flexible Web Client for R	0.9.6
DAAG	Data Analysis and Graphics Data and Functions	1.22

Write Code

R Support

Write Code Annotations:

- Navigate tabs
- Open in new window
- Save
- Find and replace
- Compile as notebook
- Run selected code
- Cursors of shared users
- Re-run previous code
- Source with or without Echo
- Show file outline
- Multiple cursors/column selection with **Alt + mouse drag**.
- Code diagnostics that appear in the margin. Hover over diagnostic symbols for details.
- Syntax highlighting based on your file's extension
- Tab completion to finish function names, file paths, arguments, and more.
- Multi-language code snippets to quickly use common blocks of code.
- Jump to function in file
- Change file type
- Working Directory
- Press **↑** to see command history
- Maximize, minimize panes
- Drag pane boundaries

R Support Annotations:

- Import data** with wizard
- History of past commands to run/copy
- Display .RPres slideshows **File > New File > R Presentation**
- Load workspace
- Save workspace
- Delete all saved objects
- Search inside environment
- Choose environment to display from list of parent environments
- Display objects as list or grid
- Displays saved objects by type with short description
- View in data viewer
- View function source code
- Create folder
- Upload file
- Delete file
- Rename file
- Path to displayed directory
- A File browser keyed to your working directory. Click on file or directory name to open.
- Change directory

Code Snippet:

```

1 # Good Start...
2
3
4
5
6 "P0030001"
7 "P0030002"
8 "P0030003"
9 "P0030004"
10
11
12 get_digit <- function() {
13   ("num" %% (10 ^ n))
14   %/% (10 ^ (n - 1))
15 }
16
17 fo
18 for {snippet}
19   foo {GlobalEnv}
20   force {base}
21
22
  
```

Console Output:

```

> foo(1)
[1] 2
> foo <- function(x) x + 1
> foo(2)
[1] 3
> foo(1)
[1] 2
  
```

<https://github.com/rstudio/cheatsheets/raw/master/rstudio-ide.pdf>

Know your R

```
>R.version
```

```
platform      _  
arch           x86_64-w64-mingw32  
os             x86_64  
system        mingw32  
status        x86_64, mingw32  
major          3  
minor         5.1  
year          2018  
month         07  
day           02  
svn rev       74947  
language       R  
version.string R version 3.5.1 (2018-07-02)  
nickname      Feather Spray
```

R Console as a Calculator

Addition and Subtraction

```
> 3+2  
[1] 5
```

```
> 3-2  
[1] 1
```

Multiplication and Division

```
> 3*2  
[1] 6
```

```
> 3/2  
[1] 1.5
```

Exponents in R

```
> 3^2  
[1] 9
```

```
> 2^3  
[1] 8
```

Constants in R

```
> pi  
[1] 3.141593
```

```
> exp(1)    base of the natural logarithm  
[1] 2.718282
```


Special values

Infinite Values

```
> Inf  
[1] Inf
```

```
> 1+Inf  
[1] Inf
```

Machine Epsilon

```
> .Machine$double.eps  
[1] 2.220446e-16
```

```
> 0>.Machine$double.eps  
[1] FALSE
```

Empty Values

```
> NULL  
NULL
```

```
> 1+NULL  
numeric(0)
```

Missing Values

```
> NA  
[1] NA
```

```
> 1+NA  
[1] NA
```

Storing and manipulating variables

Define objects `x` and `y` with values of 3 and 2, respectively:

```
> x=3
```

```
> y=2
```

Some calculations with the defined objects `x` and `y`:

```
> x+y
```

```
[1] 5
```

```
> x*y
```

```
[1] 6
```

Warning: R is case sensitive, so `x` and `X` are not the same object.

Basic R functions

Combine

```
> c(1, 3, -2)
[1] 1 3 -2
```

```
> c("a", "a", "b", "b", "a")
[1] "a" "a" "b" "b" "a"
```

Sum and Mean

```
> sum(c(1, 3, -2))
[1] 2
```

```
> mean(c(1, 3, -2))
[1] 0.6666667
```

Variance and Std. Dev.

```
> var(c(1, 3, -2))
[1] 6.333333
```

```
> sd(c(1, 3, -2))
[1] 2.516611
```

Minimum and Maximum

```
> min(c(1, 3, -2))
[1] -2
```

```
> max(c(1, 3, -2))
[1] 3
```

Basic R functions (cont.)

Define objects `x` and `y`:

```
> x=c(1,3,4,6,8)
```

```
> y=c(2,3,5,7,9)
```

Calculate the correlation:

```
> cor(x,y)
```

```
[1] 0.988765
```

Calculate the covariance:

```
> cov(x,y)
```

```
[1] 7.65
```

Combine as columns

```
> cbind(x,y)
```

	x	y
[1,]	1	2
[2,]	3	3
[3,]	4	5
[4,]	6	7
[5,]	8	9

Combine as rows

```
> rbind(x,y)
```

	[, 1]	[, 2]	[, 3]	[, 4]	[, 5]
x	1	3	4	6	8
y	2	3	5	7	9

Basic Commands

```
getwd()
setwd('E:/teaching/BayesCog/')
dir() # folders/files in the wd
ls()  # anything in the environment/workspace
print('Hello World!')
cat('Hello', 'World!')
paste0('C:/', 'Group1')
help(func)
? func # and Google!
a <- 5
a = 5
head(d) # first 6 entries
tail(d) # last 6 entries
save(varname, file = "pathname/varname.RData")
load("pathname/varname.RData")
rm(list = ls())
q()
```

RStudio - Shortcuts

Ctrl + L: clean console

Ctrl + Shift + N: create a new script

↑: command history

Ctrl(hold) + ↑: command history with certain starts

Ctrl + Enter: execute selected codes (in a script)

Editor (WIN general) - Shortcuts

Ctrl + home/Pos: go to the very top of a script

Ctrl + end/Ende: go to the very end of a script

Shift(hold) + ↑/↓: select line(s)

Ctrl(hold) + ←/→: select word(s)

Data Classes

numeric: 1.1 2.0

integer: 1 2 3

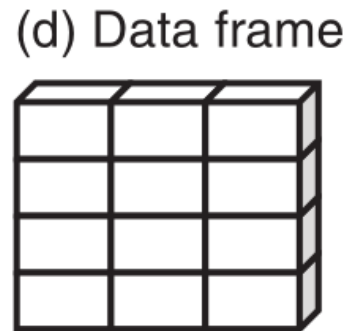
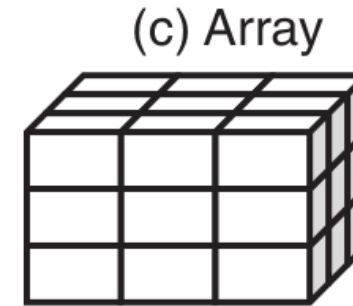
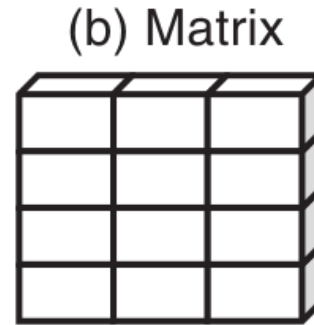
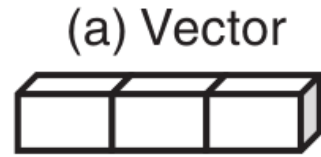
character / string: "hello world!"

logical: TRUE FALSE

factors: "male" / "female"

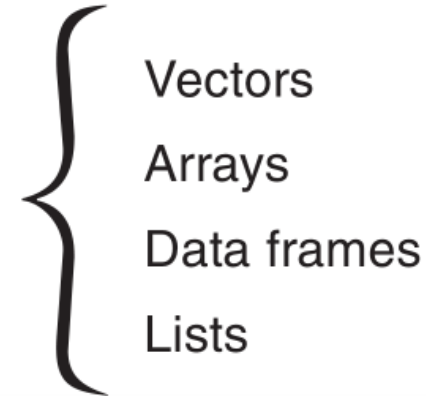
(complex: 1+2i)

Data Types



Columns can be different modes

(e) List



Exercise I

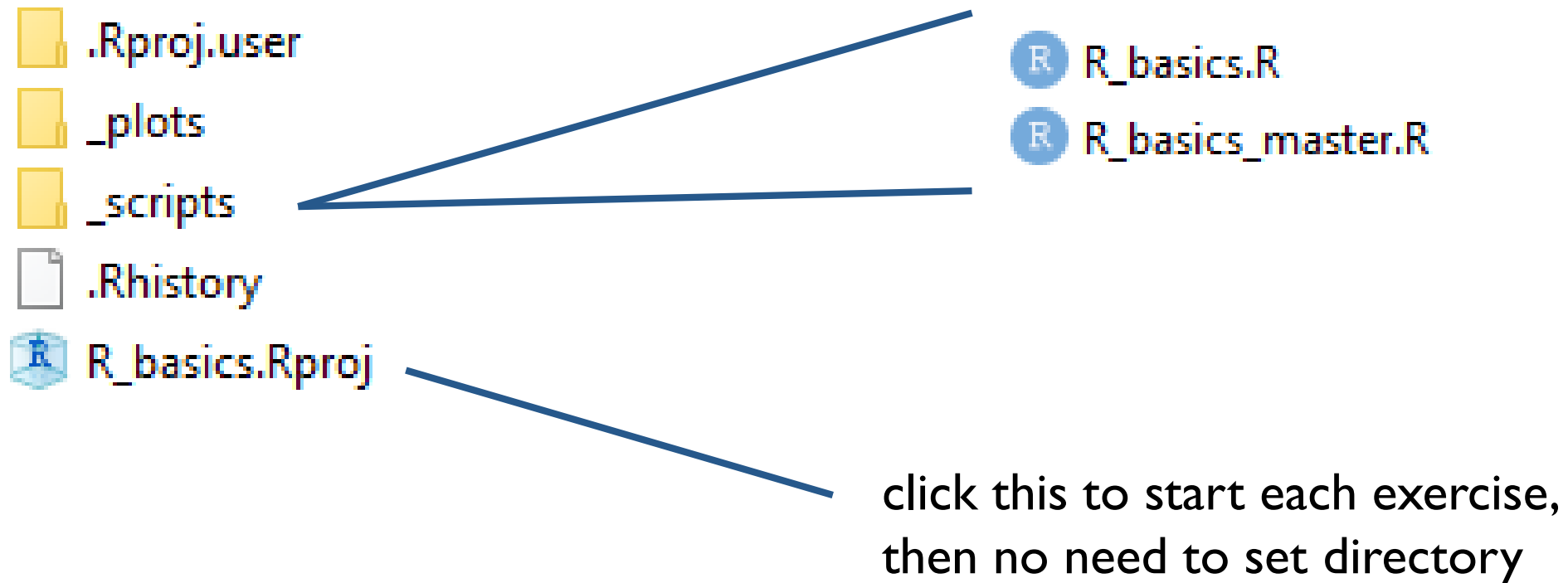
```
.../01.R_basics/_scripts/R_basics.R
```

up to “Control Flow”

TASK: practise basic R commands and data type

TIP: `class()`, `str()`

Side note: folder structure



Logical Operators

Operator	Summary
<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
==	Equal to
!=	Not equal to
!x	NOT x
x y	x OR y
x&y	x AND y

Control Flow

- if-else

```
if (cond) {  
    ..statement..  
} else {  
    ..statement..  
}
```

```
if (cond) {  
    ..statement..  
} else if (cond) {  
    ..statement..  
} else {  
    ..statement..  
}
```

- for-loop

```
for ( j in 1:n ) {  
    ..statement..  
}
```

```
for ( j in 1:J ) {  
    for ( k in 1:K ) {  
        ..statement..  
    }  
}
```

User-defined Function

```
funname <- function (input_args) {  
  .. function body ..  
  .. function body ..  
  return(output_args)  
}
```

$$sem = \sqrt{\frac{s^2}{n-1}}$$

```
sem <- function(x) {  
  sqrt( var(x,na.rm=TRUE) / (length(na.omit(x))-1) )  
}
```

Exercise II

```
.../01.R_basics/_scripts/R_basics.R
```

TASK: practise control flow and user-defined function

Packages in R

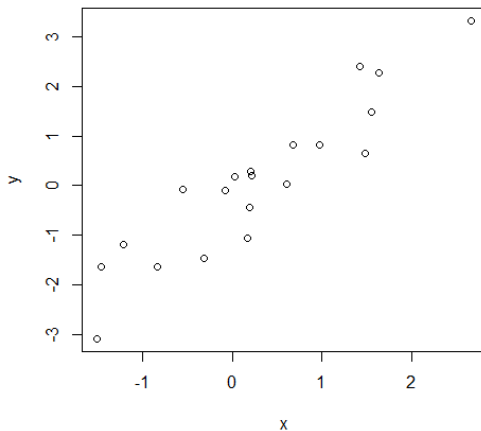
R packages are collections of functions and data sets developed by the community, to make your life a lot easier!

```
install.packages('ggplot2')  
library(ggplot2)  
detach('package:pkg')
```

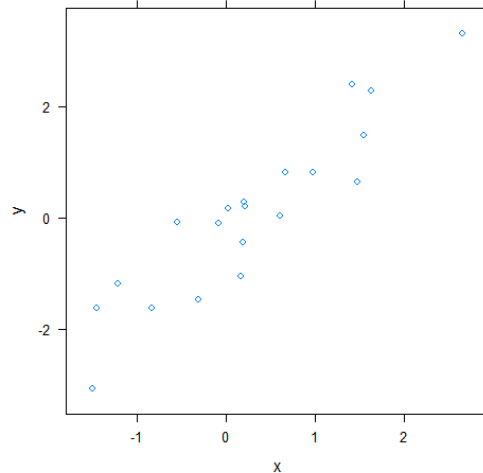

Visualization

- **built-in** plotting functions – first attempt / quick look / exploratory
- **{lattice}** – making nicer, similar to basic plotting functions
- **{ggplot2}** – making nicer, a layering philosophy

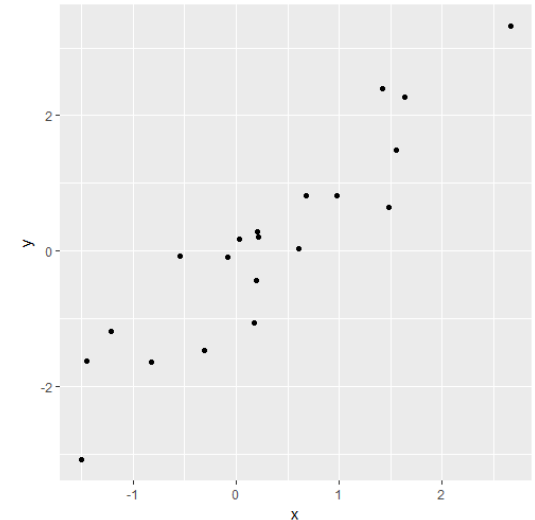
`plot(x,y)`



`lattice::xyplot(y~x)`



`ggplot2::qplot(x,y)`

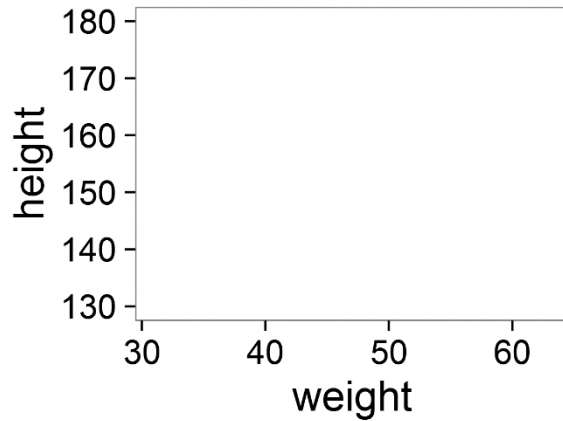


Brief Intro to ggplot2

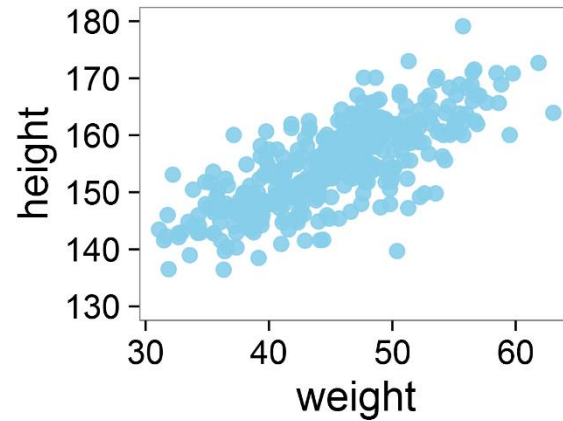
`plot` = `geometric` (points, lines, bars) + `aesthetic` (color, shape, size)

game of adding layers!

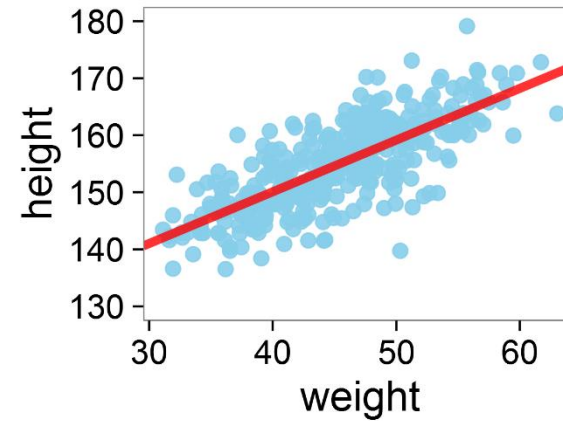
background



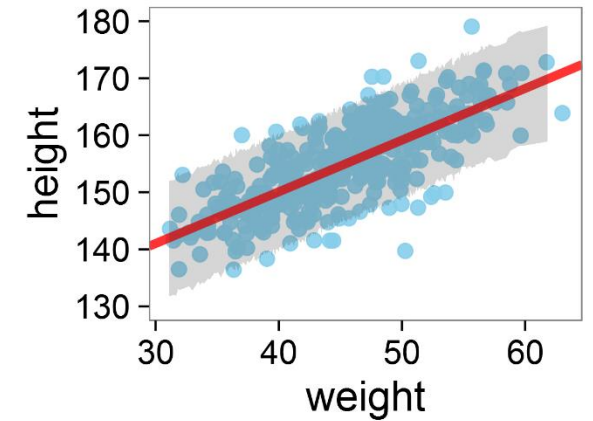
add scatters



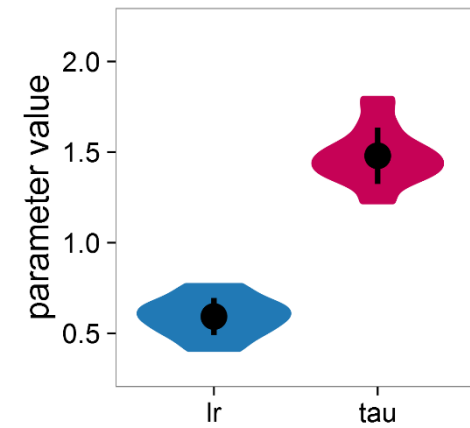
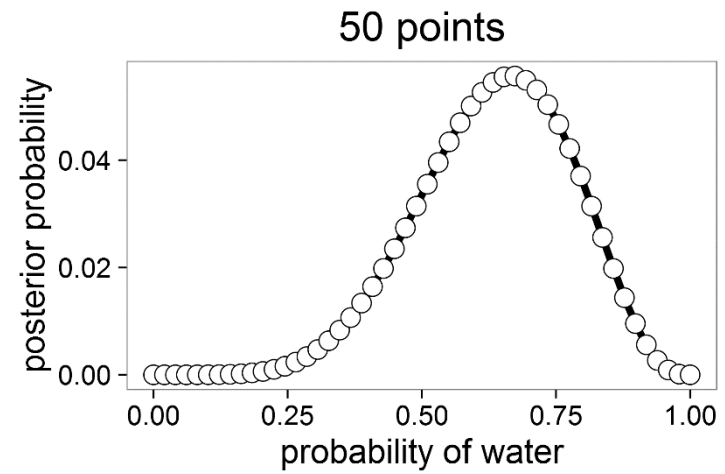
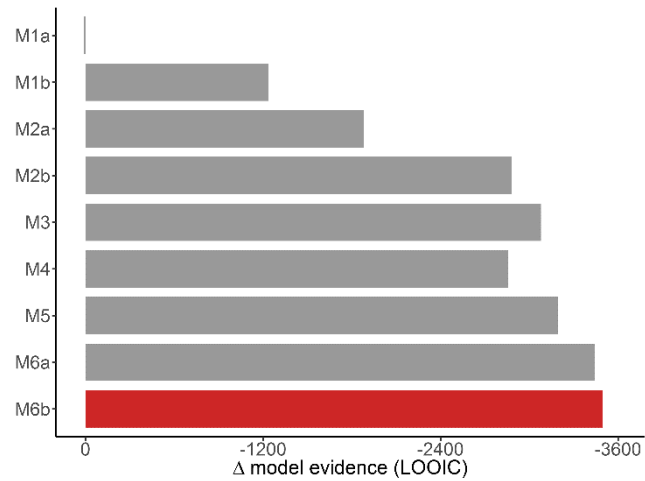
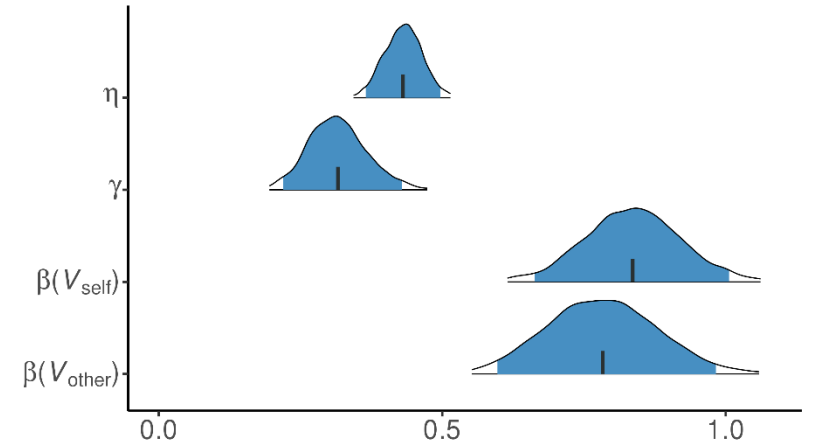
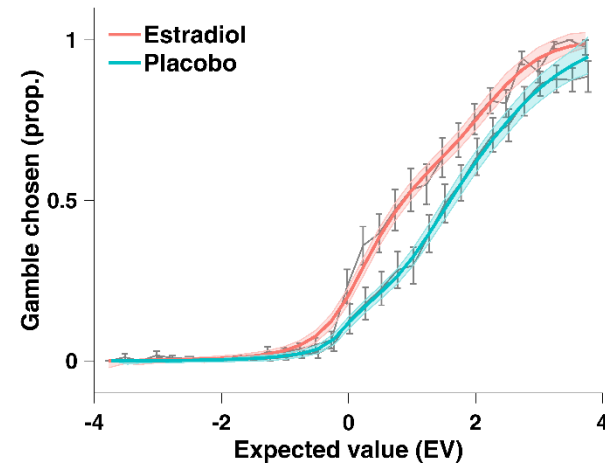
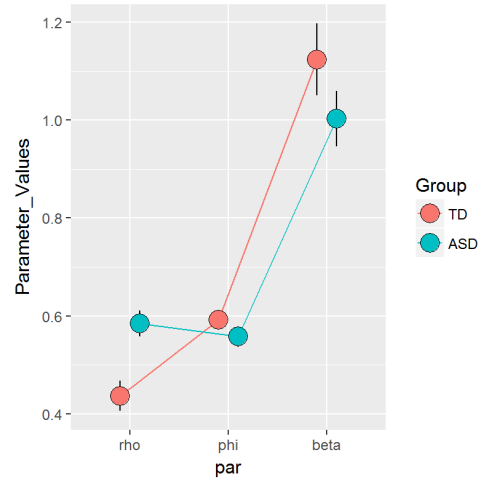
add regression line

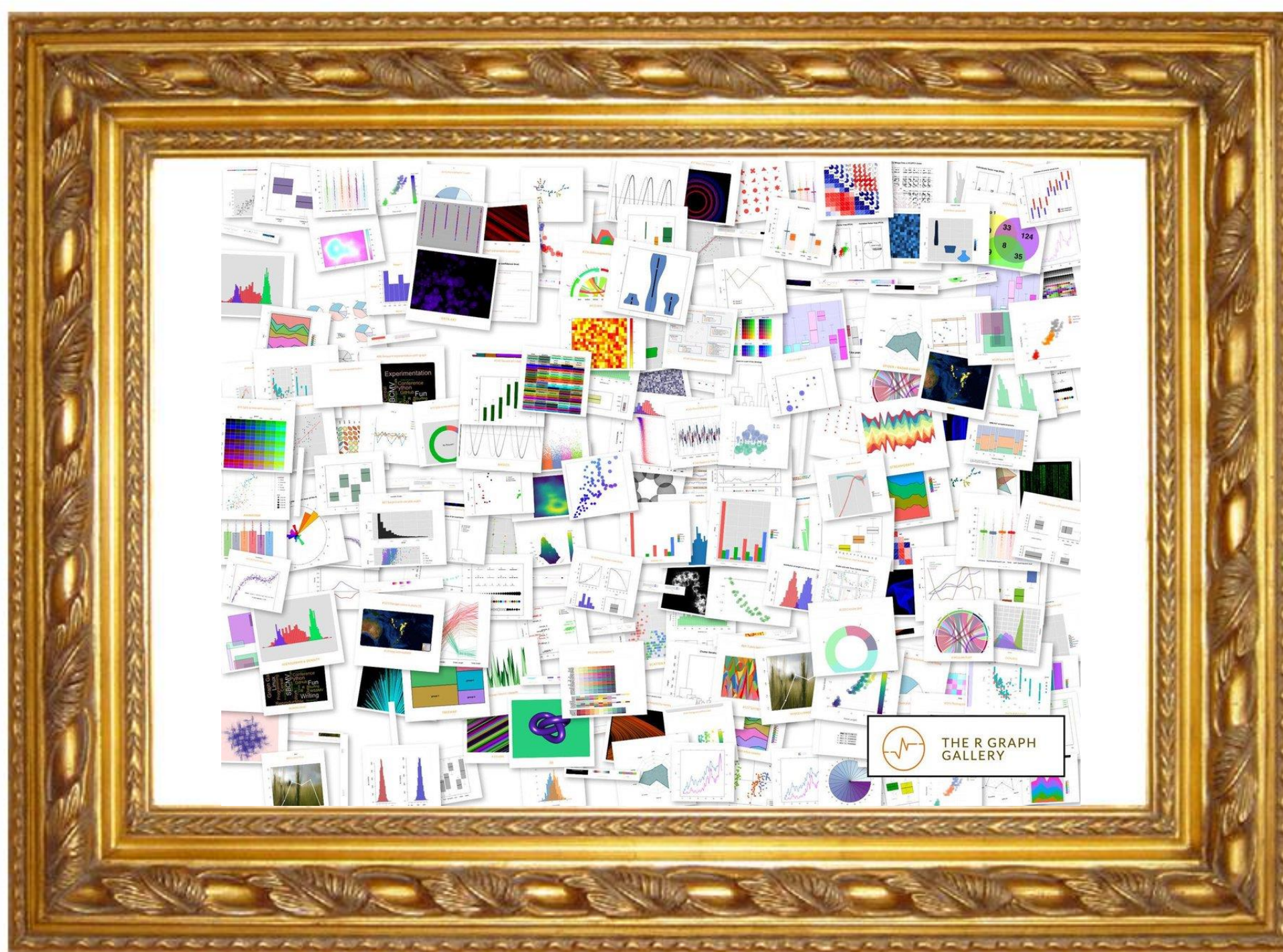


add uncertainty



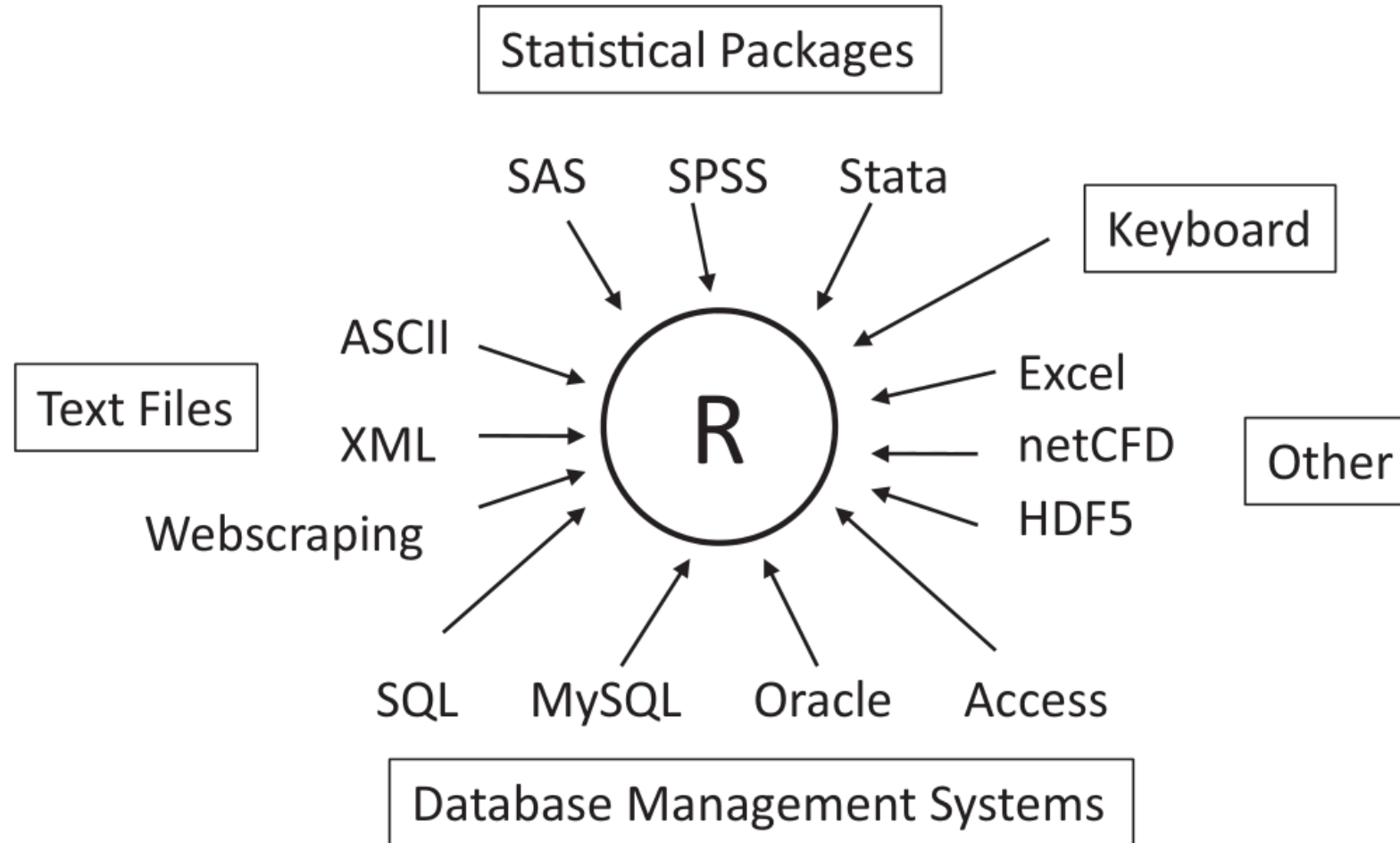
A taste of ggplot2



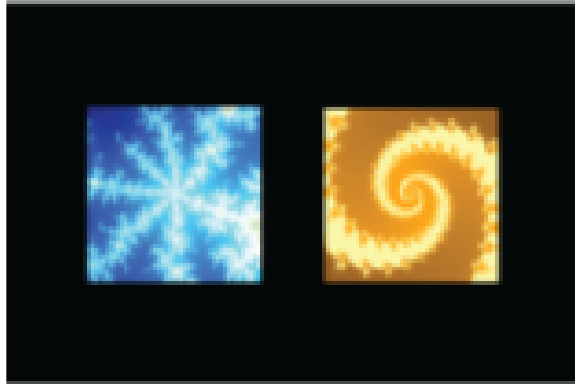


<https://www.r-graph-gallery.com/>

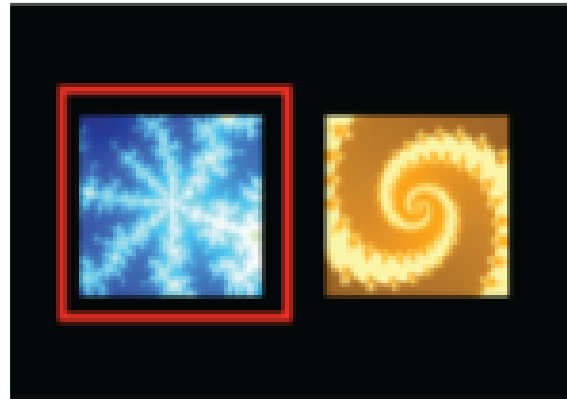
Data management



One simple experiment



choice
presentation



action
selection



outcome

reward contingency – 80:20

The data

- nSub = 10
- nTrial = 80

./_data/_raw_data/sub01/raw_data_sub01.txt

sub01	subjID, trialID, choice, outcome, correct
sub02	1,1,2,-1,1
sub03	1,2,1,1,1
sub04	1,3,1,1,1
sub05	1,4,1,1,1
sub06	1,5,2,-1,1
sub07	1,6,1,1,1
sub08	1,7,1,1,1
sub09	1,8,1,1,1
sub10	1,9,1,-1,1
	1,10,2,-1,1
	1,11,1,1,1
	1,12,1,1,1
	1,13,1,-1,2

Import some data!

```
data_dir = ('_data/RL_raw_data/sub01/raw_data_sub01.txt')  
data = read.table(data_dir, header = T, sep = ",")  
head(data)
```

	subjID	trialID	choice	outcome	correct
1	1	1	1	1	1
2	1	2	1	1	1
3	1	3	1	1	1
4	1	4	NA	1	1
5	1	5	1	-1	1
6	1	6	2	-1	1

```
sum(complete.cases(data))  
data = data[complete.cases(data),]  
dim(data[complete.cases(data),])
```


Indexing

```
data[1,1]  
data[1,]  
data[,1]  
data[1:10,]  
data[,1:2]  
data[1:10, 1:2]  
data[c(1,3,5,6), c(2,4)]  
  
data$choice
```

Exercise III

```
.../01.R_basics/_scripts/R_basics.R
```

TASK:

write a for loop

... which reads in each participant's raw data

... and reshape it in the “long format”

```
for ( j in 1:n) {  
  read.table(file, header = T, sep = ",")  
}
```

Read all the data!

```
ns = 10
data_dir = '_data/RL_raw_data'

rawdata = c();
for (s in 1:ns) {
  sub_file = file.path(data_dir, sprintf('sub%02i/raw_data_sub%02i.txt',s,s))
  sub_data = read.table(sub_file, header = T, sep = ",")
  rawdata = rbind(rawdata, sub_data)
}
rawdata = rawdata[complete.cases(rawdata),]
rawdata$accuracy = (rawdata$choice == rawdata$correct) * 1.0

acc_mean = aggregate(rawdata$accuracy, by = list(rawdata$subjID), mean)[,2]
```

Basic stats

```
mean(acc_mean)
sd(acc_mean)
sem(acc_mean)

t.test(acc_mean, mu = 0.5) # one sample t-test
```

One Sample t-test

```
data:  acc_mean
t = 13.788, df = 9, p-value = 2.34e-07
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.6962988 0.7733565
sample estimates:
mean of x
0.7348277
```

Basic correlation

```
load('_data/RL_descriptive.RData')  
descriptive$acc = acc_mean  
df = descriptive
```

```
cor.test(df$IQ, df$acc)
```

Pearson's product-moment correlation

```
data: df$IQ and df$acc
```

```
t = 4.8347, df = 8, p-value = 0.001297
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.5114810 0.9671586
```

```
sample estimates:
```

```
cor
```

```
0.8631401
```

Exercise IV

```
.../01.R_basics/_scripts/R_basics.R
```

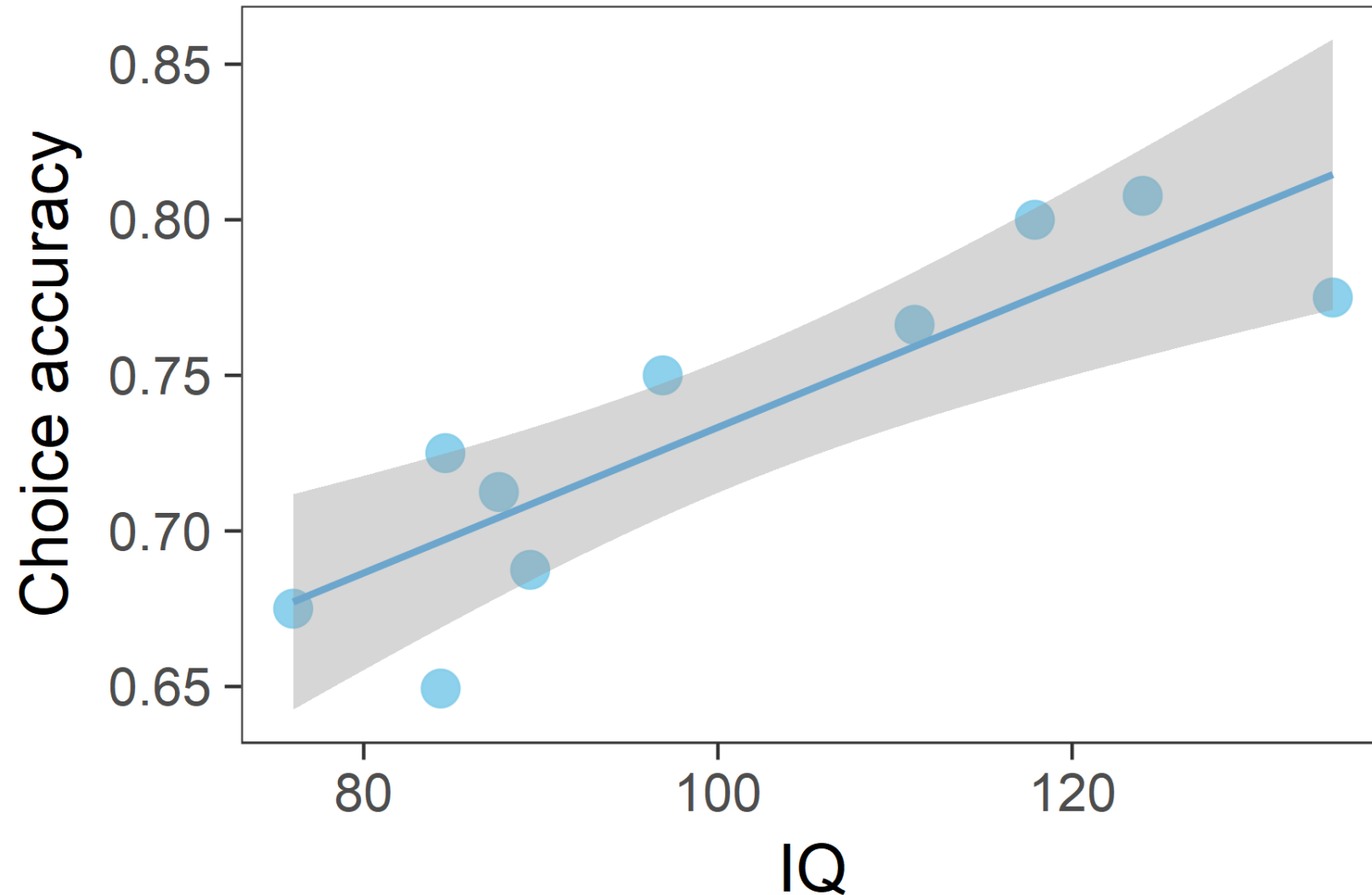
TASK:

Read in the descriptive data: `_data/descriptive.RData`
...include 'acc_mean' as a new column, and
...rename 'descriptive' as df.

Practice all the basic stats.

```
df$new_Col = new_Col
```

Plot the scatter and the regression line



Exercise V

```
.../01.R_basics/_scripts/R_basics.R
```

TASK:

Read and make sense of the ggplot functions,
... experiment make some adjustments (color marker size etc.), and
... make a similar plot for $\text{acc} \sim \text{age}$.

What is exactly the regression line in R?

```
fit1 = lm(acc ~ IQ, data = df)
summary(fit1)
```

```
Call:
lm(formula = acc ~ IQ, data = df)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.047305	-0.016277	0.007562	0.022577	0.027731

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.499292	0.049565	10.073	8.04e-06	***
IQ	0.002340	0.000484	4.835	0.0013	**

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.02885 on 8 degrees of freedom
```

```
Multiple R-squared:  0.745, Adjusted R-squared:  0.7131
```


```
F-statistic: 23.37 on 1 and 8 DF, p-value: 0.001297
```

$$\mu_i = \alpha + \beta x_i$$


$$y_i = \mu_i + \varepsilon$$

Multiple regression: more than one predictors


```
lm(target ~ predictor, data = df)
```



Target to be predicted



Variables that may help predict the target



Dataframe that contains all variables

Symbol	Example	Meaning
+	$Y \sim X1$	Include X (main effect of X)
:	$Y \sim X1:X2$	Interaction between X1 and X2
*	$Y \sim X1*X2$	Include both main affect and interaction

$Y \sim X1 + X2 + X1:X2 \Leftrightarrow Y \sim X1*X2$

Exercise VI

```
.../01.R_basics/_scripts/R_basics.R
```

TASK:

Construct the following regression models:

- main effect of age

- main effects of IQ and age

- main effects and interaction between IQ and age

use `summary()` to check R^2 and adjusted- R^2

Multiple regression: more than one predictors

```
fit1 = lm(acc ~ IQ, data = df)
fit2 = lm(acc ~ Age, data = df)
fit3 = lm(acc ~ IQ + Age, data = df)
fit4 = lm(acc ~ IQ * Age, data = df) # IQ + Age + IQ:Age
```

Model	Description	R ²	Adj-R ²	AIC
fit1	IQ only	0.75	0.71	-38.77
fit2	Age only	0.02	-0.10	-25.29
fit3	IQ Age additive	0.77	0.70	-37.76
fit4	IQ Age interactive	0.82	0.73	-38.31

Model Comparison

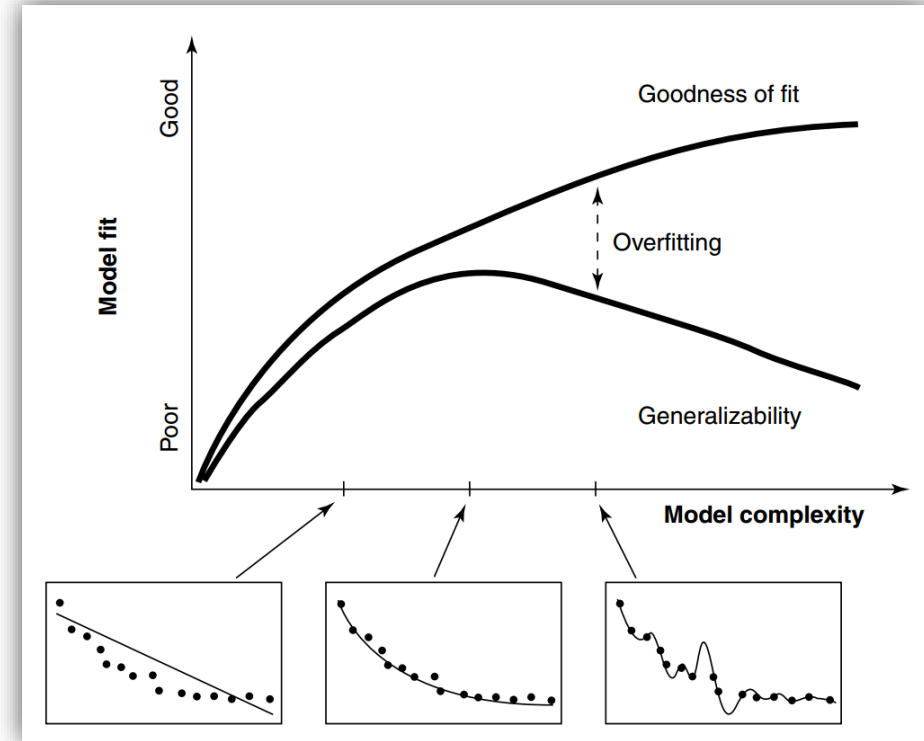
Which model provides the best **fit**?



Which model represents the best **balance** between model fit and model complexity?

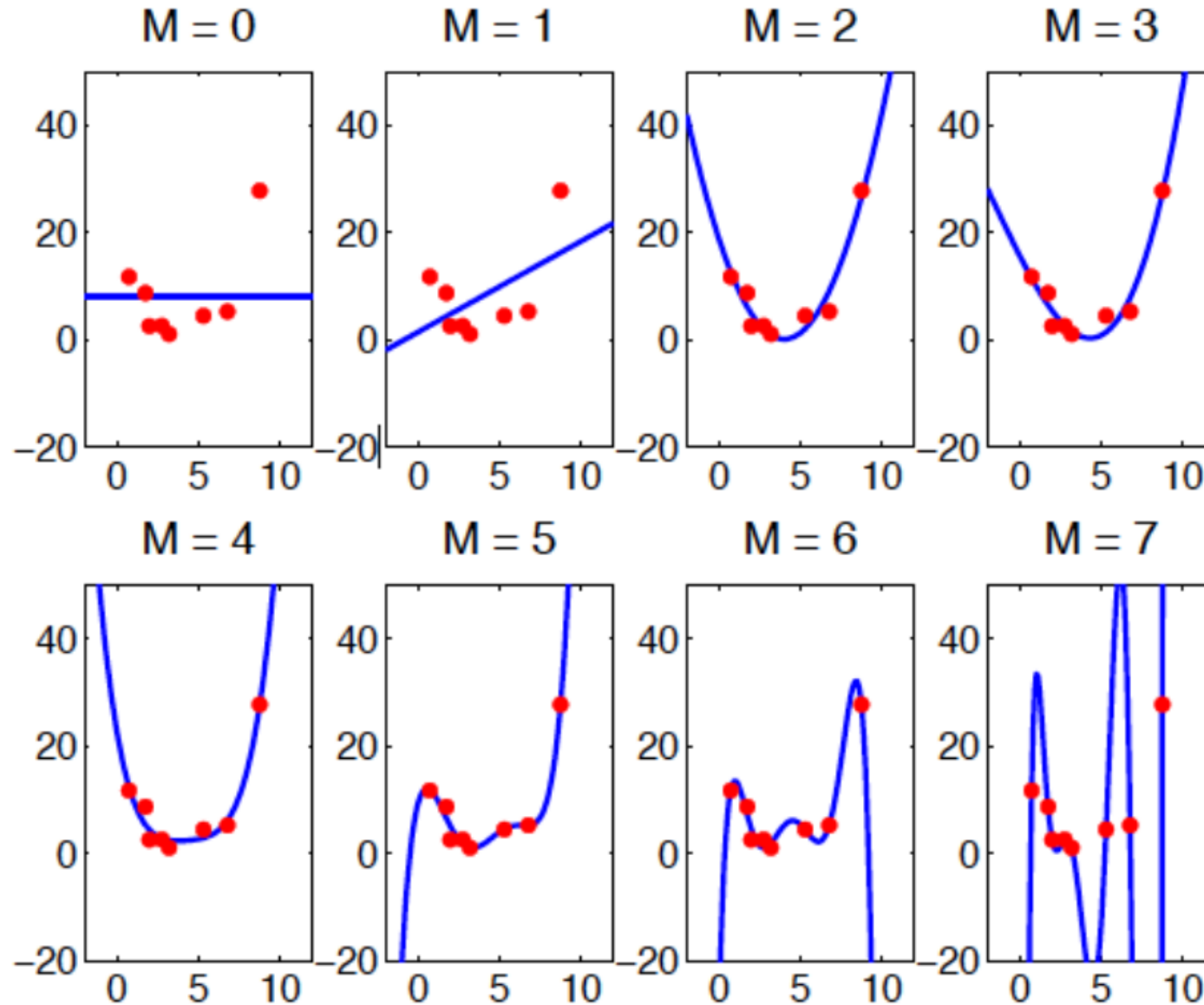
Ockham's razor:

Models with fewer assumptions are to be preferred



- overfitting: learn **too much** from the data
- underfitting: learn **too little** from the data

Which model has the highest predictive power?




Information Criteria

AIC – Akaike information criterion

DIC – Deviance Information Criterion

WAIC – Widely Applicable Information Criterion



finding the model that has
the highest out-of-sample
predictive accuracy

BIC – Bayesian Information Criterion



finding the “true” model

Compare Two Means

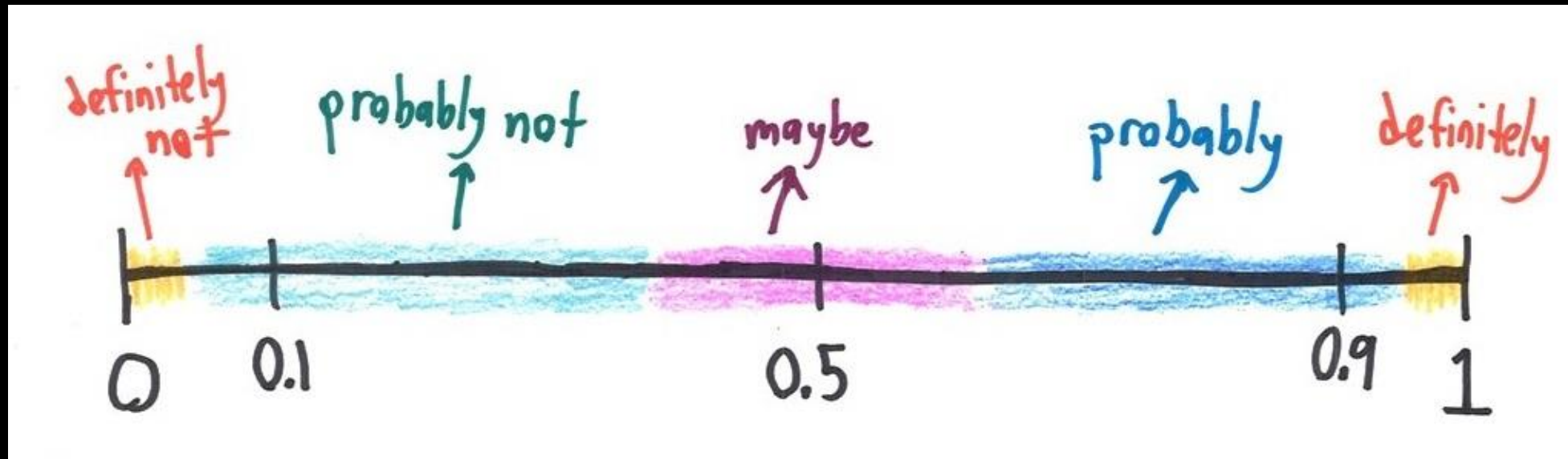
```
library(MASS)
str(UScrime)
# U1 unemployment rate of urban males 14-24.
# U2 unemployment rate of urban males 35-39.

t.test(UScrime$U1, UScrime$U2, paired=TRUE)
```

Paired t-test

```
data: UScrime$U1 and UScrime$U2
t = 32.407, df = 46, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 57.67003 65.30870
sample estimates:
mean of the differences
      61.48936
```


BASICS OF PROBABILITY



Probability

...assigning numbers to a set of possibilities

Properties (Kolmogorov, 1956)

- $p \in [0,1]$
- $\sum p = 1$
- $p(A \cup B) = p(A) + p(B)$, when A and B are *mutually exclusive*

Joint Probability and Conditional Probability

Joint Probability

$$p(A, B) = p(B, A)$$

- e.g., $p(\text{raining})$ and $p(\text{cold})$

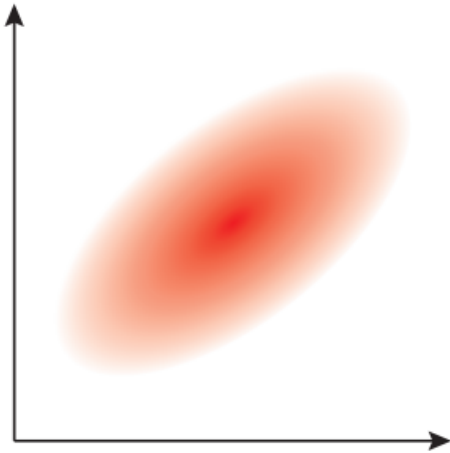
Conditional Probability

$p(A|B)$ – ‘p of A given B’

$$p(A, B) = p(A|B)p(B)$$

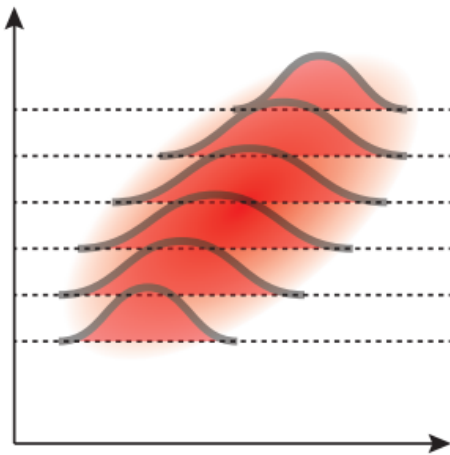
- e.g., $p(\text{raining, cold}) = p(\text{raining}|\text{cold})p(\text{cold})$

joint distribution



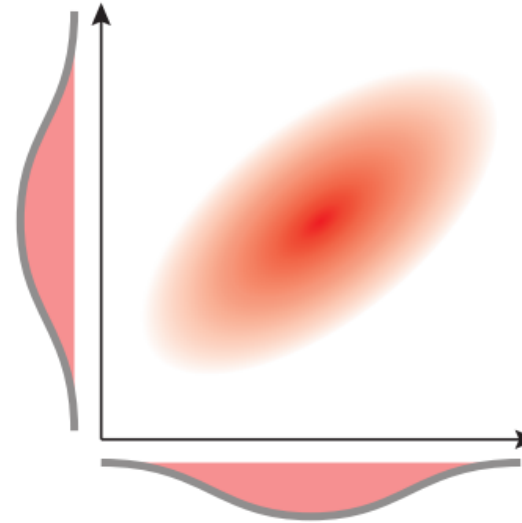
The "co-distribution" of x and y .

conditional distribution



The probability distribution of x ,
given that we know the value of y .

marginal distribution



The density of x - (or y -) values,
without knowing the other's value.



BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

$$p(A,B) = p(B,A)$$

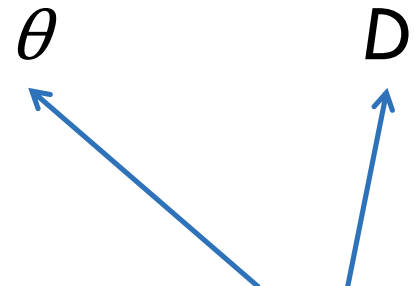
$$p(A,B) = p(A|B)p(B)$$

$$p(B,A) = p(B|A)p(A)$$

$$p(A|B)p(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Linking Data and Parameter


$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

Linking Data and Parameter

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Linking Data and Parameter

Likelihood

How plausible is the data given our parameter is true?

Prior

How plausible is our parameter before observing the data?

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Posterior

How plausible is our parameter given the observed data?

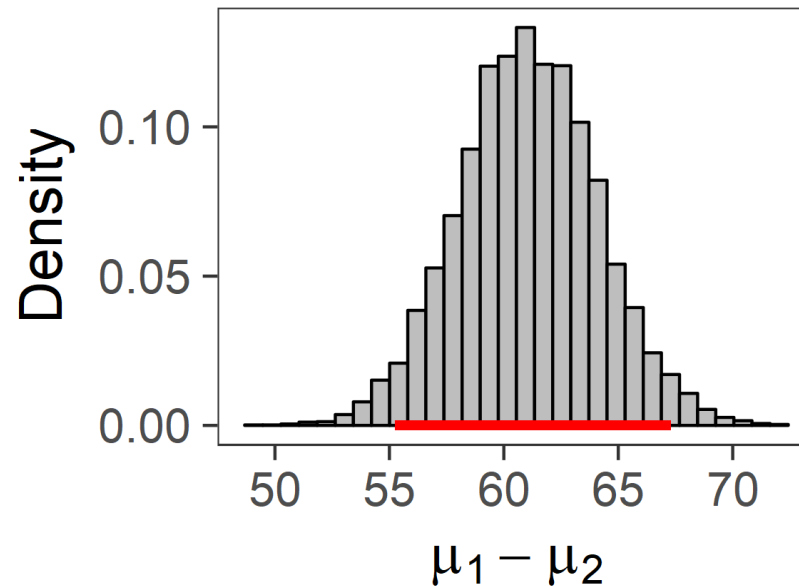
Evidence

How plausible is the data under all possible parameters?

How does that matter?

Given the data from two groups, we are interested if their means differ:

$$\rightarrow p(\mu_1 - \mu_2 | D_1, D_2)$$

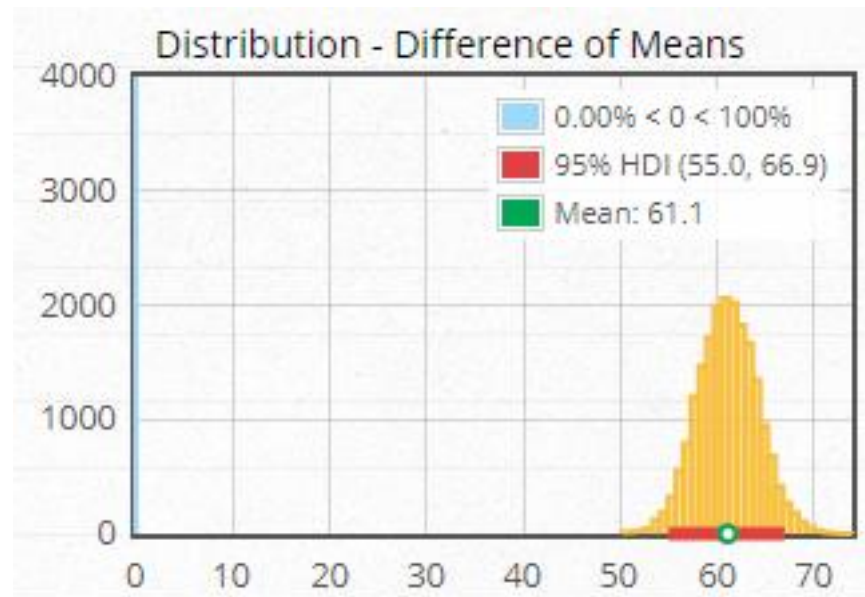


- mean: 61.06
- 95% HDI: [55.26 67.27]

Exercise VII

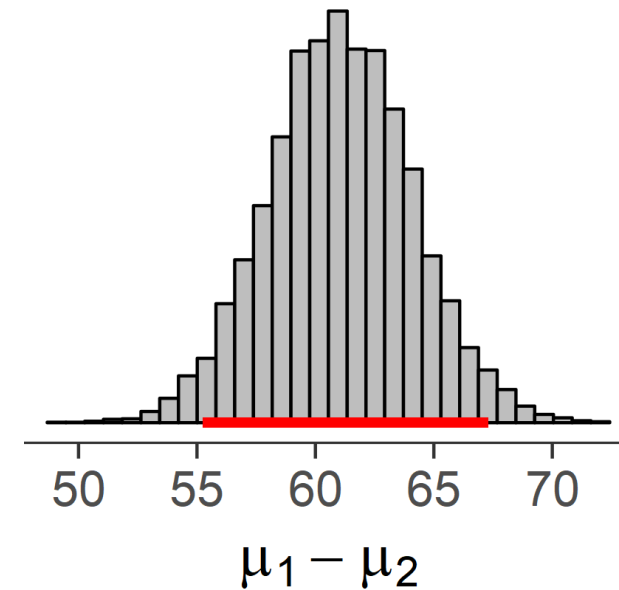
TASK:

Use the online tool to compute the posterior mean difference (U1 vs. U2) in the UScrime dataset.



Why bother?

- Incorporate prior knowledge of $(\mu_1 - \mu_2)$
- Obtain belief (uncertainty of the estimate)
- Able to accept H_0 (null hypothesis)
 - frequentist: p value is $p(D|H_0)$
- Could test more than H_1 , e.g., a bimodal distribution of the mean difference
- Have fewer assumptions



Bayes Factor

$$p(D \mid H_0) \propto p(H_0 \mid D) p(H_0)$$

$$p(D \mid H_1) \propto p(H_1 \mid D) p(H_1)$$

$$\frac{p(D \mid H_0)}{p(D \mid H_1)} = \boxed{\frac{p(H_0 \mid D)}{p(H_1 \mid D)}} \cdot \frac{p(H_0)}{p(H_1)}$$

posterior odds = Bayes factor \times prior odds

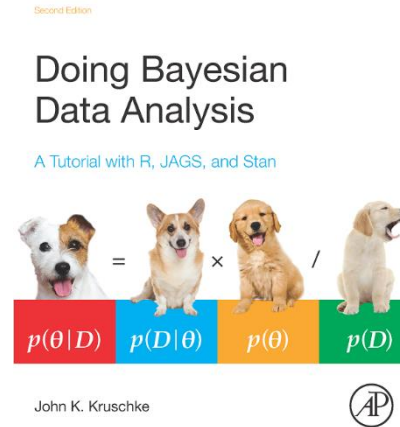
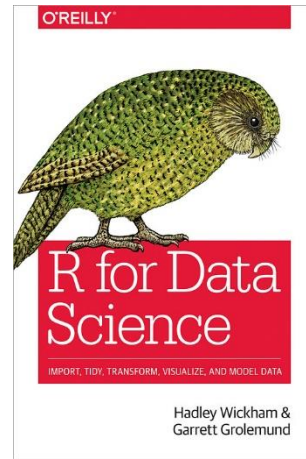
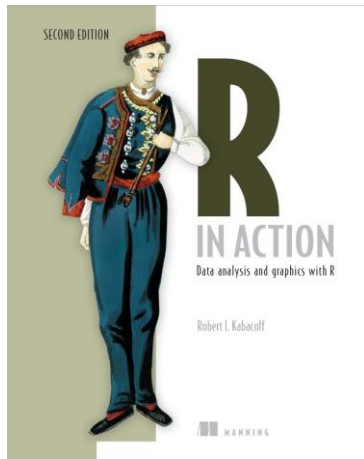
Bayes Factor

$$\text{BF} = \frac{p(H_0 | D)}{p(H_1 | D)}$$

Bayes factor	Interpretation
$B_f < 1/10$	Strong evidence for M_r
$1/10 \leq B_f < 1/3$	Moderate evidence for M_r
$1/3 \leq B_f < 1$	Weak evidence for M_r
$1 \leq B_f < 3$	Weak evidence for M_i
$3 \leq B_f < 10$	Moderate evidence for M_i
$B_f \geq 10$	Strong evidence for M_i

Source: [Min et al. \(2007\)](#).

Resources

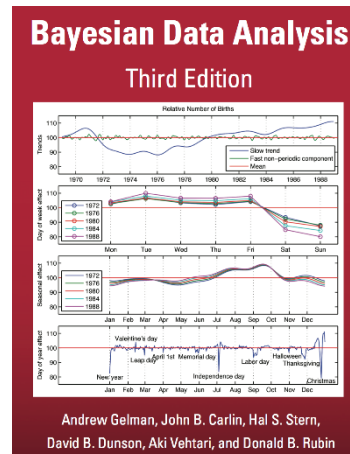
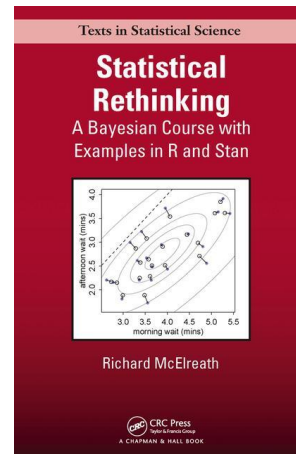
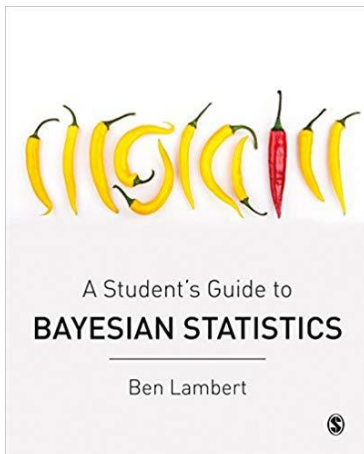


Statistical Thinking for the 21st Century

Copyright 2018 Russell A. Poldrack

Draft: 2018-11-22

<http://thinkstats.org/>



<https://jasp-stats.org/>

ANY
QUESTIONS
?

Happy R Computing!